

# Bayesian inference for Gibbs random fields using composite likelihoods

N. FRIEL<sup>1</sup>

*School of Mathematical Sciences and Complex Adaptive Systems Laboratory,  
University College Dublin, Ireland.*

July 25, 2012

## Abstract

Gibbs random fields play an important role in statistics, for example the autologistic model is commonly used to model the spatial distribution of binary variables defined on a lattice. However they are complicated to work with due to an intractability of the likelihood function. It is therefore natural to consider tractable approximations to the likelihood function. Composite likelihoods offer a principled approach to constructing such approximation. The contribution of this paper is to examine the performance of a collection of composite likelihood approximations in the context of Bayesian inference.

**Keywords and Phrases:** Composite likelihoods; Gibbs random fields; Ising model.

## 1 Introduction

Gibbs random fields play an important and varied role in statistics. The autologistic model is used to model the spatial distribution of binary random variables defined on a lattice or grid (Besag 1974). The exponential random graph model or  $p^*$  model is arguably the most popular statistical model in social network analysis (Robins *et al* 2007). Other application areas include biology, ecology and physics.

Despite their popularity, Gibbs random fields present considerable difficulties from the point of view of parameter estimation, because the likelihood function is typically intractable for all but trivially small graphs. One of the earliest approaches to overcome this difficulty is the pseudolikelihood method (Besag 1972), which approximates the joint likelihood function by the product of full-conditional distributions of all nodes. Indeed the pseudolikelihood approximation is an example of a composite likelihood, that is, a likelihood approximation consisting of a product of a joint distribution of a lower dimensional variables, each of which can be normalised. It is natural to consider approximations which refine pseudolikelihood by considering products of larger collections of variables. The purpose of this paper is to consider such composite likelihood approximations. A similar study has been conducted by Okabayshi *et al.* (2011), although from a likelihood inference perspective. As in this paper, they consider likelihood approximations consisting of a product of a joint distribution of

---

<sup>1</sup>nial.friel@ucd.ie

collections of neighbouring variables. Using the recursion method of (Reeves and Pettitt 2004) we show that larger collections of variables can be used.

This paper is organised as follows. Section 2 outlines a description of Gibbs random fields, and in particular the autologistic distribution. Composite likelihoods are introduced in Section 3. Here we focus especially on how to formulate conditional composite likelihoods for application to the autologistic model. We also focus on the issue of calibrating the composite likelihood function for use in a Bayesian context. Section 4 illustrates the performance of the various estimators for simulated data. The paper concludes with some remarks in Section 5.

## 2 Discrete-valued Markov random fields

Discrete Markov random fields play an important role in several areas of statistics including spatial statistics and social network analysis. The autologistic model, popularised by Besag (1972) which has the Ising model as a special case, is widely used in analysis of binary spatial data defined on a lattice. The exponential random graph (or  $p^*$ ) model is frequently used to model relational network data. See (Robins *et al* 2007) for an excellent introduction to this body of work.

Let  $y = \{y_1, \dots, y_n\}$  denote realised data defined on a set of nodes  $\{1, \dots, n\}$  of a graph, where each observation  $y_i$  takes values from some finite state space. The likelihood of  $y$  given a vector of parameters  $\theta = (\theta_1, \dots, \theta_m)$  is defined as

$$f(y|\theta) \propto \exp(\theta^T s(y)) := q(y|\theta), \quad (1)$$

where  $s(y) = (s_1(y), \dots, s_m(y))$  is a vector of sufficient statistics. The constant of proportionality in (1),

$$z(\theta) = \sum_{y \in Y} \exp(\theta^T s(y)),$$

depends on the parameters  $\theta$ , and is a summation over all possible realisation of the Gibbs random field. Clearly,  $z(\theta)$  is intractable for all but trivially small situations. This poses serious difficulties in terms of estimating the parameter vector  $\theta$ .

One of the earliest approaches to overcome the intractability of the (1) is the pseudolikelihood method (Besag 1975) which approximates the joint distribution of  $y$  as the product of full-conditional distributions for each  $y_i$ ,

$$f_{pseudo}(y) = \prod_{i=1}^n f(y_i|y_{-i}, \theta),$$

where  $y_{-i}$  denotes  $y \setminus \{y_i\}$ . This approximation has been shown to lead to unreliable estimates of  $\theta$ , for example, (Rydén and Titterton 1998), (Friel *et al* 2009). This is in fact one of the earliest composite likelihood approximations, and we will outline work in this area further in Section 3. Note also that Monte Carlo approaches have also been exploited to estimate the intractable likelihood, for example the Monte Carlo maximum likelihood estimator of Geyer

and Thompson (1992). More recently, auxiliary variable approaches have been presented to tackle this problem through the single auxiliary variable method (Møller *et al* 2006) and the exchange algorithm (Murray *et al* 2006).

The autologistic model, first proposed by Besag (1972), is defined on a regular lattice of size  $m \times m'$ , where  $n = mm'$ . It is used to model the spatial distribution of binary variables, taking values  $-1, 1$ . The autologistic model is defined in terms of two sufficient statistics,

$$s_0(y) = \sum_{i=1}^n y_i, \quad s_1(y) = \sum_{j=1}^n \sum_{i \sim j} y_i y_j,$$

where the notation  $i \sim j$  means that lattice point  $i$  is a neighbour of lattice point  $j$ . Henceforth we assume that the lattice points have been indexed from top to bottom in each column and where columns are ordered from left to right. For example, for a first order neighbourhood model where an interior point  $y_i$  has neighbours  $\{y_{i-m}, y_{i-1}, y_{i+1}, y_{i+m}\}$ . Along the edges of the lattice each point has either 2 or 3 neighbours. The full-conditional of  $y_i$  can be written as

$$p(y_i | y_{-i}, \theta) \propto \exp(\theta_0 y_i + \theta_1 y_i (y_{i-m} + y_{i-1} + y_{i+1} + y_{i+m})), \quad (2)$$

where  $y_{-i}$  denotes  $y$  excluding  $y_i$ . As before, the conditional distribution is modified along the edges of the lattice. The Hammersley-Clifford theorem (Besag 1974) shows the equivalence between the model defined in (2) and in (1). Note that parameter  $\theta_0$  controls the relative abundance of  $-1$  and  $+1$  values. The parameter  $\theta_1$  controls the level of spatial aggregation. When  $\theta > 0$ , neighbouring values tend to take similar values, thereby yielding homogeneous regions in the lattice. Note that the Ising model is a special case, resulting from  $\theta_0 = 0$ .

### 3 Composite likelihoods

There has been considerable recent interest in composite likelihood methods in the statistics literature under such headings as pairwise likelihood methods (Nott and Rydén 1999), composite likelihoods (Heagerty and Lele 1998), (Cox and Reid 2004) and split-data likelihoods (Rydén 1994). These concepts have long-standing antecedents such as Besag's pseudolikelihood (Besag 1975). The basic idea is to work with a likelihood made up of factors, each of which corresponds to the joint probability function of a small number of variables, two in the case of pairwise likelihoods.

Let us now return to the case where  $y$  is a realisation from an autologistic distribution. Following the previous section we denote  $A = \{1, \dots, mm'\}$  as an index set for the lattice points. Following (Asuncion *et al* 2010) we consider a general form of composite likelihood written as

$$f(y|\theta) \approx \prod_{i=1}^C p(y_{A_i} | y_{B_i}, \theta) := CL(y|\theta). \quad (3)$$

Some special cases arise:

1.  $A_i = A$ ,  $B_i = \emptyset$ ,  $C = 1$  corresponds to the full likelihood.

2.  $B_i = \emptyset$  is often termed *marginal composite likelihood*.
3.  $B_i = A \setminus A_i$  is often termed *conditional composite likelihood*.

The focus of this paper is on conditional composite likelihoods. Note that the pseudolikelihood approximation is a special of 3. where each  $A_i$  is a singleton. We restrict each  $A_i$  to be of the same dimension and in particular to correspond to contiguous square 'blocks' of lattice points of size  $k \times k$ . In terms of the value of  $C$  in case 3., an exhaustive set of blocks would result in  $C = (m - k + 1) \times (n - k + 1)$ . In particular, we allow the collection of blocks  $\{A_i\}$  to overlap with one another. Finally, it is worth noting that in our context marginal distributions,  $p(y_{A_i}|\theta)$ , for index sets  $A_i$ , are rarely, if ever, available. Hence we don't consider marginal composite likelihoods in this context.

Our interest here is to compare the performance of estimation of  $\theta$  using the conditional composite likelihood and especially to understand the statistical efficiency as the block size  $k$  increases. It should also be evident that the computational complexity of this approximation will increase dramatically with  $k$ , the size of the blocks. Consequently our interest here is to explore the trade-off between using a larger block size, with a smaller number of blocks  $C$  in (3).

### 3.1 Computing full-conditional distributions of $A_i$

The conditional composite likelihood which we described above relies on evaluating

$$p(y_{A_i}|y_{-A_i}, \theta) = \frac{\exp(\theta_0 s_0(y_{A_i}) + s_1(y_{A_i}|y_{-A_i}))}{z(\theta, y_{A_i})}, \quad (4)$$

where

$$s_0(y_{A_i}) = \sum_{j \in A_i} y_j, \quad s_1(y_{A_i}|y_{-A_i}) = \sum_{j \in A_i} \sum_{l \sim j} y_l y_j.$$

Also the normalising constant now includes the argument  $y_{A_i}$  emphasising that it involves a summation over all possible realisations of sub-lattices defined on the set  $A_i$  and conditioned on the realised  $y_{-A_i}$ . First we describe an approach to compute the overall normalising constant for a lattice, without any conditioning on a boundary.

Generalised recursions for computing the normalizing constant of general factorisable models such as the autologistic models have been proposed by Reeves and Pettitt (2004), generalising a result known for hidden Markov Models (e.g. Zucchini and Guttorp 1991; Scott 2002). This method applies to autologistic lattices with a small number of rows, up to about 20, and is based on an algebraic simplification due to the reduction in dependence arising from the Markov property. It applies to un-normalized likelihoods that can be expressed as a product of factors, each of which is dependent on only a subset of the lattice sites. We can write  $q(y|\theta)$  in factorisable form as

$$q(y|\theta) = \prod_{i=1}^n q_i(\mathbf{y}_i|\beta),$$

where each factor  $q_i$  depends on a subset  $\mathbf{y}_i = y_i, y_{i+1}, \dots, y_{i+m}$  of  $y$ , where  $m$  is defined to be the *lag* of the model. We may define each factor as

$$q_i(\mathbf{y}_i, \beta) = \exp\{\theta_0 y_i + \theta_1 y_i (y_{i+1} + y_{i+m})\} \quad (5)$$

for all  $i$ , except when  $i$  corresponds to a lattice point on the last row or last column, in which case  $y_{i+1}$  or  $y_{i+m}$ , respectively, drops out of (5).

As a result of this factorisation, the summation for the normalizing constant,

$$z(\theta) = \sum_y \prod_{i=1}^n q_i(\mathbf{y}_i | \theta)$$

can be represented as

$$z(\theta) = \sum_{y_n} q_n(\mathbf{y}_n | \theta) \sum_{y_{n-1}} q_{n-1}(\mathbf{y}_{n-1} | \theta) \cdots \sum_{y_1} q_1(\mathbf{y}_1 | \theta) \quad (6)$$

which can be computed much more efficiently than the straightforward summation over the  $2^n$  possible lattice realisations. Full details of a recursive algorithm to compute the above can be found in Reeves and Pettitt (2004). Note that this algorithm was extended in (Friel and Rue 2007) to also allow exact draws from  $p(y|\theta)$

The minimum lag representation for an autologistic lattice with a first order neighbourhood occurs for  $r$  given by the smaller of the number of rows or columns in the lattice. Identifying the number of rows with the smaller dimension of the lattice, the computation time increases by a factor of two for each additional row, but linearly for additional columns. It is straightforward to extend this algorithm to allow one to compute the normalising constant in (4), so that the summation is over the variables  $y_{A_i}$  and each factor involves conditioning on the set  $y_{-A_i}$ .

### 3.2 Bayesian composite likelihoods

The focus of interest in Bayesian inference is the posterior distribution

$$p(\theta|y) \propto f(y|\theta)p(\theta).$$

Our proposal here is to replace the true likelihood  $f(y|\theta)$  with a conditional composite likelihood approximation, leading us to focus on the approximated posterior distribution

$$p^*(\theta|y) \propto CL(y|\theta)p(\theta).$$

Surprisingly, there is very little literature on the use of composite likelihoods in the Bayesian setting, although Pauli *et al.* (2011) present a discussion on the use of conditional composite likelihoods. Indeed this paper suggests, following (Lindsay 1988), that a composite likelihood should take the general form

$$f(y|\theta) \approx \prod_{i=1}^C p(y_{A_i} | y_{B_i}, \theta)^{w_i}, \quad (7)$$

where  $w_i$  are positive weights. In all experiments carried out here, we assume that  $w_i = 1$ , and empirically we observe that non-calibrated composite likelihood leads to an approximated posterior distribution with substantially lower variability than the true posterior distribution, leading to overly precise precision about posterior parameters.

## 4 Examples

Here we simulated 20 realisations from a first-order Ising model all defined on a  $16 \times 16$  lattice, with a single interaction parameter  $\theta = 0.4$ , which is close to the critical phase transition beyond which all realised lattices takes either value  $+1$  or  $-1$ . This parameter setting is the most challenging for the Ising model, since realised lattices exhibit strong spatial correlation around this parameter value. For a lattice of this dimension it is possible to exactly calculate the normalising constant  $z(\theta)$ . Using a fine grid of  $\{\theta_i\}$  values, the right hand side of:

$$p(\theta_i|y) \propto \frac{q(y|\theta_i)}{z(\theta_i)} p(\theta_i), \quad i = 1, \dots, n.$$

can be evaluated exactly. Summing up the right hand side yields an estimate of the evidence,  $p(y)$ , which is the constant of normalisation for the expression above and which in turn can be used to give a very precise estimate of  $p(\theta|y)$ . This serves as a ground truth against which to compare with the posterior estimates of  $\theta$  using the various composite likelihood estimators. Exact calculation of  $z(\theta)$  and the approach described above to get a precise estimate of the evidence relies on algorithms developed in (Friel and Rue 2007). For this experiment, we choose a uniform  $[-10, 10]$  prior for  $\theta$ .

In terms of MCMC implementation, 5000 iterations were used with a burn in period of 1,000 iterations for each dataset. Computation was carried out on a desktop PC with a 3.33Ghz processor and with 4Gb of memory. Computation time for each of the different composite likelihoods was approximately constant and took 0.004 second of CPU time per iteration. This was achieved by exhaustively using all  $3 \times 3$  blocks in the  $\text{CCL}_3$  approximation, 40% of all  $4 \times 4$  blocks, 20% of all  $5 \times 5$  blocks and 10% of all  $6 \times 6$  blocks in the  $\text{CCL}_4$ ,  $\text{CCL}_5$  and  $\text{CCL}_6$  approximations, respectively. The results for simulations involving the  $16 \times 16$  lattices are displayed in Figure 1. Here each of the conditional composite likelihood methods perform better than pseudolikelihood. Each of the composite likelihood approximations performed equally well, although the  $\text{CC}_4$ ,  $\text{CCL}_5$  and  $\text{CCL}_6$  display larger spread than  $\text{CCL}_3$ .

It is apparent from Table 1 that the estimated posterior variance of  $\theta$  for each of the approximations are generally lower than the true posterior variances. In fact the conditional composite likelihood approximations lead to posterior variance estimates that are smaller by a factor of 10. This strongly suggests that the conditional composite likelihoods need to be calibrated in some form.

In a similar vein to the previous experiment, here we examined the performance of the various approximations on simulated data defined on larger  $50 \times 50$  lattice. In this instance we can't analytically compute the true likelihood, however here we used the exchange algorithm (Murray, Ghahramani and MacKay 2006) to generate draws from the target posterior,

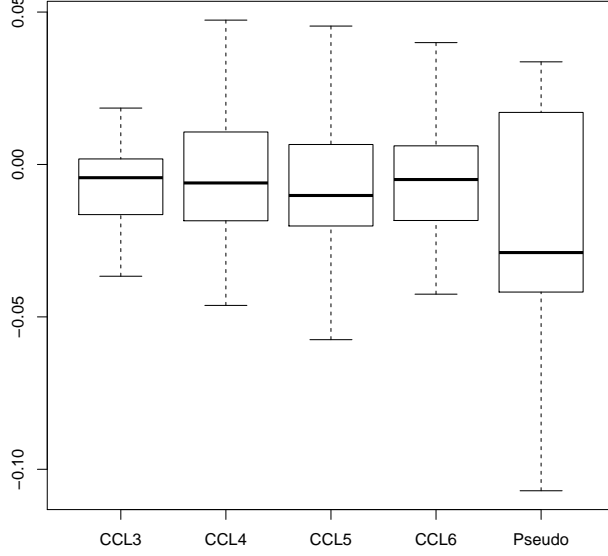


Figure 1:  $16 \times 16$  lattices: Boxplot displaying the bias of the estimate of  $\theta$  for 20 datasets corresponding to each of 4 conditional composite likelihood approximations,  $CCL_3$ ,  $CCL_4$ ,  $CCL_5$  and  $CCL_6$  (corresponding to block size  $3 \times 3$ ,  $4 \times 4$ ,  $5 \times 5$ ,  $6 \times 6$ , respectively) and also pseudolikelihood estimator. Note that the computational time for each of the composite likelihood approximations was held constant.

$CCL_3$	$CCL_4$	$CCL_5$	$CCL_6$	Pseudo	True
$2.1 \times 10^{-4}$	$3.0 \times 10^{-4}$	$3.3 \times 10^{-4}$	$4.3 \times 10^{-4}$	$2.1 \times 10^{-3}$	$1.6 \times 10^{-3}$

Table 1:  $16 \times 16$ : Average posterior variance for  $\theta$  for each of 20 datasets.

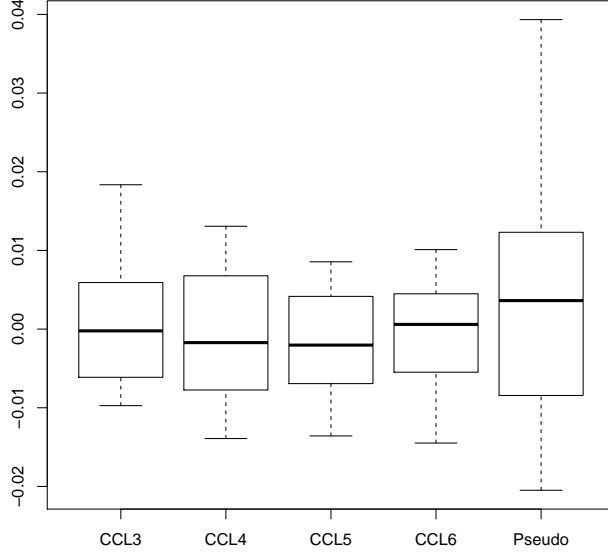


Figure 2:  $50 \times 50$  lattices: Boxplot displaying the bias of the estimate of  $\theta$  for 20 datasets corresponding to each of 4 conditional composite likelihood approximations,  $CCL_3$ ,  $CCL_4$ ,  $CCL_5$  and  $CCL_6$  (corresponding to block size  $3 \times 3$ ,  $4 \times 4$ ,  $5 \times 5$ ,  $6 \times 6$ , respectively) and also pseudolikelihood estimator. Note that the computational time for each of the composite likelihood estimator was held constant.

from a very long MCMC run. The simulation study was otherwise similar in every other respect. Computation time was approximately 0.1 second per iteration of the MCMC algorithm using the various composite likelihood approximations. Here the performance of the conditional composite likelihood was again similar to each other, and better generally, in terms of lower bias, than posterior mean estimation using the pseudolikelihood approximation. Here, similar to the previous experiment, we see in Table 2 that the posterior variance based on the various the conditional composite likelihood are considerably small, than that estimated by the exchange algorithm.

$CCL_3$	$CCL_4$	$CCL_5$	$CCL_6$	Pseudo	True
$1.5 \times 10^{-3}$	$3.3 \times 10^{-5}$	$2.2 \times 10^{-5}$	$2.6 \times 10^{-5}$	$2.1 \times 10^{-4}$	$1.3 \times 10^{-2}$

Table 2:  $50 \times 50$ : Average posterior variance for  $\theta$  for each of 20 datasets.

#### 4.1 Why is the posterior variance of estimators based on composite likelihoods overly precise?

The results of this section suggest that using conditional composite likelihoods leads to considerably underestimated posterior variances. A possible explanation for this behaviour may be due to a type of 'annealing' effect where the true likelihood is replaced by a powered



version of it, leading to an overly concentrated likelihood function. Here the true likelihood  $f(y|\theta)$  is replaced by  $\prod_{i=1}^C p(y_{A_i}|y_{A \setminus A_i}, \theta)^{w_i}$ . Suppose that  $w_i = 1$  for all  $i$  (as is the case in all of the experiments carried out here) and suppose further that  $C$  is large, whereby potentially many blocks overlap. In this scenario the set of interactions in the true likelihood will be a subset of all the interactions in the conditional composite likelihood, due to the overlapping blocks, and this will in turn lead to an annealing of the true likelihood function.

## 5 Conclusion

This paper has illustrated the important role that composite likelihood approximations can play in the statistical analysis of Gibbs random fields, and in particular in the Ising and autologistic models in spatial statistics. This paper has focused on the use of conditional composite likelihoods, based on tractable full-conditional distributions over blocks of lattices points and shows much promise. However it is evident that the posterior distribution of the interaction parameter  $\theta$  is too concentrated and therefore underestimates the posterior variance. An important research question is to ask how to correctly calibrate the conditional composite likelihood so that it achieves the correct variance.

The computational complexity of this approximation increases dramatically as the size of blocks increases and our study here shows that efficient parameter estimation can result by considering conditional composite likelihoods based on a subset of possible blocks, thereby reducing computation time. For future work it would be interesting to understand how composite likelihoods can be usefully employed for more challenging Markov random fields models.

**Acknowledgements:** Nial Friel’s research was supported by a Science Foundation Ireland Research Frontiers Program grant, 09/RFP/MTH2199.

## References

- Asuncion, A. U., Q. Liu, A. T. Ihler and P. Smyth (2010), Learning with Blocks: Composite Likelihood and Contrastive Divergence. *AISTATS, Journal of Machine Learning Research: W&CP* **9**, 33–40
- Besag, J. E. (1972), Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B* **34**, 75–83
- Besag, J. E. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 192–236
- Besag, J. E. (1975), Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195
- Cox, D. R. and N. Reid (2004), A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737

- Friel, N., A. N. Pettitt, R. Reeves and E. Wit (2009), Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics* **18**, 243–261
- Friel, N. and H. Rue (2007), Recursive computing and simulation-free inference for general factorizable models. *Biometrika* **94**, 661–672
- Geyer, C. J. and E. A. Thompson (1992), Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society, Series B* **54**, 657–699
- Heagerty, P. J. and S. R. Lele (1998), A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**, 1099–1111
- Lindsay, B. (1988), *Statistical inference from Stochastic processes*, vol. 80, chap. Composite likelihoods, pp. 221–239. American Mathematical Society, Providence, RI
- Møller, J., A. N. Pettitt, R. Reeves and K. K. Berthelsen (2006), An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–458
- Murray, I., Z. Ghahramani and D. MacKay (2006), MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, AUAI Press
- Nott, D. J. and T. Rydén (1999), Pairwise likelihood methods for inference in image models. *Biometrika* **86**, 661–676
- Okabayshi, S., L. Johnson and C. J. Geyer (2011), Extending pseudo-likelihood for Potts models. *Statistica Sinica* pp. 331–347
- Pauli, F., W. Racugno and L. Ventura (2011), Bayesian composite marginal likelihoods. *Statistica Sinica* pp. 149–164
- Reeves, R. and A. N. Pettitt (2004), Efficient recursions for general factorisable models. *Biometrika* **91**, 751–757
- Robins, G., P. Pattison, Y. Kalish and D. Lusher (2007), An introduction to exponential random graph models for social networks. *Social Networks* **29**(2), 169–348
- Rydén, T. (1994), Consistent and asymptotically normal parameter estimates for hidden Markov models. *The Annals of Statistics* **22**, 1884–1895
- Rydén, T. and D. M. Titterton (1998), Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics* **7**, 194–211
- Scott, S. L. (2002), Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Am. Statist. Assoc.* **97**(457), 337–51
- Zucchini, W. and P. Guttorp (1991), A hidden Markov model for space time precipitation. *Water Resour. Res.* **27**, 1917–23